DOCUMENT RESUME

ED 383 719 TM 023 140

AUTHOR Enright, Mary K.; Powers, Donald E.
TITLE Validating the GRE Analytical Ability Measure against Faculty Ratings of Analytical Reasoning Skills. GRE Board Professional Report No. 86-06P.
INSTITUTION Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-90-22
PUB DATE Jan 91
NOTE 61p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Academic Achievement; *College Faculty; Criteria; Grades (Scholastic); *Graduate Students; Graduate Study; Higher Education; Inferences; Rating Scales; Student Evaluation; *Test Construction; *Thinking Skills
IDENTIFIERS *Analytic Ability; *Graduate Record Examinations

ABSTRACT

The aim of this study was to develop a criterion of graduate school success as an alternative to first-year average. Faculty rating scales of students' analytical abilities were developed as a potential criterion against which to validate both the current Graduate Record Examinations (GRE) analytical measure and its future modifications. The instrument included six separate scales for faculty to rate individual students on: (1) analyzing arguments; (2) drawing inferences; (3) defining problems; (4) reasoning inductively; (5) generating alternatives; and (6) overall analytical style. The scales were completed by faculty members in 24 departments representing 6 disciplines. Over all departments, 132 faculty members rated 623 students, 58% of whom were rated by at least 3 raters. Faculty raters were not able to distinguish among students on the six individual scales. In addition, evidence was found that the ratings and first-year grades measure somewhat different aspects of success in graduate school. Results were also mixed with respect to the validity of faculty ratings of students' analytical abilities. Faculty ratings of students' analytical skills appear to have been influenced by students' verbal reasoning skills. It is recommended that the development of these scales be continued. Nine tables and two figures present study findings. An appendix contains the rating scales. (Contains 33 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

GRE® RESEARCH

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H.I. Braun
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 383 719

Validating the GRE Analytical Ability Measure Against Faculty Ratings of Analytical Reasoning Skills

Mary K. Enright and Donald E. Powers

January 1991

GRE Board Professional Report No. 86-06P
ETS Research Report 90-22

BEST COPY AVAILABLE

TM023140

Educational Testing Service, Princeton, New Jersey

Validating the GRE Analytical Ability Measure Against Faculty Ratings of Analytical Reasoning Skills

Mary K. Enright and Donald E. Powers

GRE Board Report No. 86-06P

January 1991

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Graduate Record Examinations and Educational Testing Service are U.S. registered trademarks of Educational Testing Service; GRE, ETS, and the ETS logo design are registered in the U.S.A. and in many other countries.

Copyright © 1991 by Educational Testing Service. All rights reserved.

Abstract

The aim of this study was to develop a criterion of graduate school success as an alternative to first-year average. More specifically, faculty rating scales of students' analytical abilities were developed as a potential criterion against which to validate both the current Graduate Record Examinations (GRE) analytical measure and future modifications of it.

The rating instrument was based on previous research (Powers & Enright, 1986, 1987), which identified a number of independent dimensions underlying faculty perceptions of the importance of a wide variety of reasoning skills. The instrument included six separate scales for faculty to rate individual students with respect to their skills in analyzing arguments, drawing inferences, defining problems, reasoning inductively, and generating alternatives, as well as their overall analytical style.

The rating scales were completed by faculty members in a sample of 24 graduate departments representing six disciplines. Three important results have implications for the use of faculty ratings as a criterion of success. First, faculty raters were not able to distinguish among students on the six individual scales, which exhibited very high intercorrelations. This suggests that the rating instrument could be simplified for future use.

Secondly, although the ratings and first-year grades were highly correlated, indicating that both criteria reflect success in graduate school, evidence that ratings and first-year grades measure somewhat different aspects of success in graduate school was also found. Each of the three GRE General Test measures--verbal, quantitative, and analytical--was more highly correlated, on the average, with ratings than with first-year averages. Undergraduate grades, on the other hand, correlated better with first-year grades than with the ratings.

Finally, results were mixed with respect to the validity of faculty ratings of students' analytical abilities. When the three GRE measures were ranked with respect to their predictive effectiveness for each department, the analytical measure was significantly more often the best or second best predictor of faculty ratings than of first-year average, while the verbal and quantitative measures tended to be the best predictors about equally often for ratings and grades. This suggests that the ratings may be more reflective of analytical ability than of verbal or quantitative ability. However, the verbal measure was, on average, more highly correlated with faculty ratings of students' analytical skills than was the analytical measure. This suggests that faculty ratings of students' analytical skills may have been influenced by students' verbal reasoning skills. This failure to find unequivocal evidence of discriminant validity of the ratings may reflect problems with the ratings, with the way in which faculty rated students, or with the discriminant validity of the analytical measure. A recommendation was made to continue research on the development of these scales.

Validating the GRE Analytical Ability Measure Against Faculty Ratings of Analytical Reasoning Skills

One significant undertaking in admissions testing in recent years has been the GRE Board's effort to extend graduate admissions testing beyond strictly verbal and quantitative domains, thereby facilitating a broader definition of academic talent. This effort resulted eventually in a measure of analytical ability, which was introduced in the GRE General Test on an experimental basis in 1977.

Research revealed that the new measure reflected a dimension that was distinguishable from the verbal and quantitative abilities measured by the test (Powers & Swinton, 1981), and that analytical scores exhibited moderate relationships with performance in graduate school (Wilson, 1982). Unfortunately, two of the four analytical item types proved to be susceptible both to special test preparation (Swinton & Powers, 1983; Powers & Swinton, 1984) and to within-test practice (Swinton, Wild, & Wallmark, 1983). The deletion of these two problematic item types in 1981 yielded a revised measure that was heavily weighted toward deductive reasoning and thus much more narrowly focused than the original measure.

For the current version of the analytical measure, evidence for both discriminant and convergent validity has been mixed (Stricker & Rock, 1985; Wilson, 1985). Specifically, the two remaining item types appear generally to have less in common with each other than with either the verbal or the quantitative items used in the test, most likely because all three measures of the General Test are designed to tap reasoning ability. And, although a distinct, but relatively weak analytical factor has been detected in several academic disciplines, it has been defined largely by only one of the two item types in the measure (Schaeffer & Kingston, 1988). In addition, correlations of analytical scores with first-year graduate grades have been modest, with analytical scores adding relatively little to the prediction of grades beyond the contributions made by verbal and quantitative scores (Kingston, 1985).

Finally, the analytical measure has exhibited some peculiar properties. Not only does it seem to behave differently in different fields, but it appears to have a chameleon-like nature, exhibiting validities that are similar to those of the verbal measure in relatively verbal fields and similar to those of the quantitative measure in quantitatively oriented fields of study (Wilson, 1982). Because of these phenomena, research on the development and evaluation of additional analytical item types is underway with a view toward improving the current analytical measure.

The predictive validity of the analytical measure has been assessed primarily on the basis of its relationship to first-year graduate grades, with all of their well-known problems as a criterion of success in graduate school. Hartnett & Willingham (1979) and Wild, Swinton, and Brown (in preparation) have provided thorough discussions of the role of the criterion in test validation efforts and of the problems associated with using first-year grades as a criterion of success in graduate school. Most importantly, the range of first-year grades is greatly restricted: graduate faculty typically assign grades of A or B, thus limiting the size of the correlation that can be obtained between test scores and grades.

Hartnett and Willingham (1979) noted the potential of faculty ratings as an alternative criterion of success in graduate school, and data collected through the GRE Validity Study Service (VSS) indicate that such ratings can be a useful criterion (Burton & Turner, 1983). In fact, faculty ratings are the most commonly used optional criterion among departments that participate in the VSS. Indeed, for these departments, GRE scores predict faculty ratings better than they predict first-year grades. However, rating instruments are known to have their own shortcomings, including restriction of range and halo effects (Saal, Downey, & Lahey, 1980). Unfortunately, the data submitted to the VSS are not extensive enough to permit the comparative evaluation of first-year grades and faculty ratings as criteria of success in graduate school.

The present study involved the development and evaluation of faculty ratings as alternative criteria of success. The specific focus was on ratings of the analytical skills or abilities that have been suggested by graduate faculty as being most important for successful graduate study in their fields (Powers & Enright, 1987). The objectives were to develop and evaluate a criterion that reflects faculty judgments of analytical skills involved in successful graduate study, and to explore the validation of the current GRE analytical measure against this criterion. More important, the availability of a suitable standardized criterion was envisioned as having potential for evaluating progress toward an improved measure of analytical ability.

Background

Recently, Powers and Enright (1987) asked graduate faculty in six fields of study (chemistry, computer science, education, engineering, English, and psychology) to judge:

(a) the importance for academic success of a wide variety of analytical, reasoning, or thinking skills (e.g., the capacity to identify the assumptions on which an argument is based), particularly as these skills differentiate successful from marginal graduate students

(b) the criticality of specific incidents related to thinking or reasoning that may have caused faculty to either raise or lower their estimation of a student's analytical ability (e.g., failing to qualify a conclusion as appropriate)

(c) the seriousness of various reasoning or thinking "flaws" that faculty may have observed in their students (e.g., confusing correlation with causation)

The ratings of 96 skills, incidents, and "flaws," culled from the literature and a preliminary survey of faculty, were reduced through factor

analysis to five dimensions for reasoning skills and to three for critical incidents. Two additional dimensions were found to underlie reasoning "flaws," but these were not easily interpreted and therefore were not considered in developing rating scales for this project. The various dimensions, some of which were more prominent than others, are given in Table 1.

These dimensions were judged by faculty to be required in differing degrees in particular fields of study. For example, English faculty rated the items defining the Explanation factor as more important than any others for successful study in English. Computer science faculty, on the other hand, judged this dimension to be far less important than others, such as Problem Analysis (which English faculty rated as the least important of all). It suffices to say that faculty perceptions of the importance of each dimension differed dramatically by field of study according to the requirements for mastery in each discipline.

Method

Instrument Development

As one of the most ubiquitous forms of evaluation, rating scales have been the subject of a great deal of research, which has involved investigating such aspects of rating systems as the characteristics of raters and ratees, the nature of both the rating instrument and the rating process, and the context in which the ratings are made (including the purpose of the ratings) (Landy & Farr, 1980). The findings from this research guided the development of the rating instrument used in this study. More specifically, the issues considered in designing our instrument included what type of scale to use, how to assess raters' familiarity with ratees, and what kind of instructions to provide faculty raters.

Type of scale. For several reasons, standard graphic scales (which may differ according to the particular verbal or numerical labels used to anchor scales) were used in the present study rather than rankings (which may involve paired comparisons or forced distributions of ratees), or checklists (which may involve series of adjectives or descriptive statements to be checked and scored by summing over all items).

In preliminary discussions, graduate faculty suggested that a complete ranking of all graduate students in a department would involve false precision and might be resisted by faculty members. Checklists, on the other hand, did not seem to lend themselves particularly well to rating the kinds of analytical abilities in which we were interested, nor did they appear to be a particularly effective way to build upon the knowledge base that was available, i.e., the faculty perceptions that were gathered previously. Finally, behaviorally anchored scales, while gaining some prominence in

industrial/organizational settings in previous years, did not appear to meet our needs completely either. In an earlier application of this kind of rating, Carlson, Reilly, Mahoney, and Casserly (1976) encountered considerable resistance by graduate faculty to completing behaviorally anchored scales, apparently because the use of specific behavioral examples frustrated faculty who had not had ample opportunity to observe those behaviors. Other users of behaviorally anchored scales have encountered other problems--for example, in identifying appropriate anchors for particular portions of the scale (Harari & Zedeck, 1973; Landy & Guion, 1970). Aside from these specific difficulties, a major objection to using behaviorally anchored ratings has been the considerable cost involved in developing, testing, and revising these scales (Landy & Farr, 1980).

Anastasi (1979) and Smith (1976) have summarized the results of research on ratings and have concluded generally that who does the rating makes more difference than do the particular characteristics of the rating scale or the specific rating techniques used. Landy and Farr (1980), for instance, concluded that despite much good research there is still, for the amount of development involved, no efficient and psychometrically sound alternative to the traditional graphic rating scale. Given these considerations and our previous experience in collecting faculty perceptions of analytical skills, a standard graphic scale appeared to offer the most promise for obtaining faculty ratings of students' analytical skills.

A rating instrument consisting of seven scales was constructed (Appendix A). The first six scales were designed to assess the kinds of reasoning skills that emerged from the factor analysis of faculty ratings of the importance of various reasoning skills and are described in Table 2. In addition, a seventh scale was included to obtain an indication of each rater's familiarity with each student, since familiarity is likely to affect the quality of ratings and therefore the relationship of ratings to test performance (Freeberg, 1969; Landy & Farr, 1980). Moreover, the importance of obtaining some indication of raters' acquaintance with ratees has been reinforced recently in the new Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985), which state that when criteria are composed of rater judgments, the degree of knowledge that raters have concerning ratee performance should be reported.

Five-point response scales were used in the instrument because research on the optimal number of scale categories tends to indicate that little is gained in going beyond five categories (e.g., Jenkins & Taber, 1977; Landy & Farr, 1980; Lissitz & Green, 1975). For the six reasoning scales, response categories were differentiated in terms of the degree to which a particular student was more or less able than other students on each of these dimensions. In addition, there was a response category to indicate that the faculty member could not rate a student with respect to a particular skill.

The instrument was assembled physically in a way thought to minimize halo error. Faculty were asked to rate all students on one scale before moving to the next one. As ratings were completed for a particular scale, a

page was turned so the rater could not readily see the ratings that had been made on scales completed previously. (This, of course, did not necessarily prevent raters from remembering the ratings they had made earlier.)

Instructions to raters. Research on ratings also shows that well-trained raters are more likely to provide reliable and valid ratings than are untrained ones (see review by Landy & Farr, 1980). In addition, the characteristics of ratings often depend on the purpose for which they are obtained. For example, ratings obtained for administrative purposes tend to be more lenient than those obtained for research purposes. Moreover, as Carlson et al. (1976) suggested, the graduate faculty in their study would have been more likely to cooperate in providing ratings if the goal of the project had been to validate GRE scores, instead of to develop criteria for administrative purposes.

The most practical way to provide relevant training in this study was through the directions that accompanied the rating form. Therefore, the instructions emphasized the following points:

- the "research only" nature of the project
- the importance of trying to rate students independently on each scale
- the importance of using the full 5-point range of the scale
- encouragement to remember or observe both positive and negative instances of performance

In addition, raters were encouraged to review the rating instrument and to observe students for a week or more before making their ratings.

Sample Selection

A number of different sources were used to identify graduate departments to participate in this study. Initially, the 252 graduate departments contacted for the prior study were approached (Powers & Enright, 1987). This sample, obtained from the data tapes of the Higher Education General Information Survey (NCES, 1984), consisted of 42 departments in each of six fields: chemistry, computer science, education, engineering, English, and psychology. For each field, a sample was drawn of 64 graduate programs that, according to the Directory of Graduate Programs (GRE/CGS, 1983), either required or recommended GRE scores. In this manner, 40 institutions were selected for the final sample for each field. In addition, one institution having a relatively large proportion of Black students and one having a relatively large number of Hispanic students were included in the samples for each field, thus increasing the total number of institutions to 42 per field. This sample was augmented by other graduate departments in these fields that

ranked high in terms of the number of GRE score reports that were sent to them or that were participating in the GRE Validity Service. Altogether, a total of 370 departments was contacted. About 11% of these departments indicated interest in participating after receiving a detailed description of the study, and about 60% of this "interested" sample eventually submitted usable data. In the final sample of 24 departments, the number per field ranged from 2 in chemistry to 7 in psychology.

Securing departmental and faculty cooperation proved more difficult than expected. In the previous study (Powers & Enright, 1987), 65% of the departments contacted indicated interest in participating as compared to 11% in the present study. A number of factors contributed to this decreased interest and varied by field. For all fields, the fact that graduate faculty were extremely busy was a factor. In addition, there were concerns about students' rights to privacy and the legality of releasing information about individual students. An example of a field-specific obstacle was the organization of many chemistry departments into laboratory groups. Some chemistry departments indicated that students were known only by a single faculty member, with whom they worked almost exclusively, and not by other faculty.

Procedures

Most departments were first sent a letter explaining that studies were going to be conducted on the development of faculty rating scales, which were to be used as research instruments to validate the Graduate Record Examinations. These departments were asked to indicate their interest by completing and returning a form including information about their graduate programs. For some departments the letter was followed by a telephone call to explore interest and to ascertain the characteristics of the department. A $300 honorarium was offered to departments for participating.

A more detailed letter about the study procedures was sent to departments that expressed interest and had sufficient numbers of students. This letter was followed by a phone call to confirm a department's ability and willingness to participate. In the latter stages of our research this procedure was modified so that departments were sent only the second, more detailed letter, and the honorarium ($75) was paid to the individual who coordinated the data collection for the department. Individual faculty members received $25 each for completing the rating scales.

Participating departments were sent instructions requesting that three to five faculty members in each department rate between 10 and 30 first-year students on the scale and that a similar number of faculty rate 10 or more post-first-year students on the assumption that faculty knowledge of post first-year students could differ from their knowledge of first-year students. We asked that only faculty who had some relevant contact with the students serve as raters. Raters of the two groups of students did not have to be the same individual faculty members. The definition of a post-first year student

varied with departments. For master's programs, it usually meant second-year students, while for doctoral programs it referred to third- or fourth-year students. We also asked departments to provide students' GRE scores, their undergraduate grade point averages (UGPA), and their first-year graduate school averages (FYA). Ratings were collected near the end of the academic year or in the first semester of a subsequent year. That is, all students who were rated had been graduate students for nearly a full year or more.

Results

Three central issues were addressed by the data analysis. The first concerned the quality of the ratings data. The second involved the comparison between first-year grades and faculty ratings of students' analytical abilities as a criterion of success in graduate school. The third issue concerned whether evidence for the discriminant validity of the analytical measure would be found in relation to the ratings of analytical skills.

Sample Characteristics

Table 3 summarizes by discipline the mean GRE scores, UGPA, and FYA for the students rated in this study. Students in the physical sciences (chemistry, computer science, engineering) had relatively high scores on the GRE quantitative measure and low scores on the verbal measure. Conversely, students in English had high mean GRE verbal scores and comparatively low quantitative scores. In the field of psychology, mean scores were moderately high on all the GRE measures while, in education, mean scores were relatively low on all GRE measures. These differences among the disciplines in GRE score patterns are consistent with differences evident in data summarizing the performance of students applying to graduate school (ETS, 1988) and among departments that have participated in the GRE Validity Study Service (Burton & Turner, 1983).

Analysis of the Rating Data

Participating departments varied in the degree to which they were able to comply with our request that at least 3 faculty members rate 10 students in common. Over all departments, a total of 132 faculty members served as raters and 623 students were rated. Of these students, 19% were rated by only one rater, 23% were rated by 2 raters, and 58% were rated by 3 or more raters. The number of pairs of raters that rated at least 5 students in common was 145.

Preliminary analysis of the rating data focused on factors such as leniency, restriction of range, correlations among the six scales, and reliability. The mean rating on each scale and a mean rating over all scales

is presented by discipline in Table 4. (An inspection of means for first-year and post-first-year students and of correlations of scales with other variables for each of these groups revealed no major differences. Thus, all subsequent analyses were based on samples pooled across educational level.)

The rating data and graduate FYA data can be contrasted with respect to leniency and restriction of range, as measured by the standard deviation. In theory, grades in most graduate programs represent a 5-point rating scale (0-4). However, in practice most graduate faculty assign primarily A's, B's, and an occasional C. In comparing mean FYA in Table 3 with the overall mean rating in Table 4, we see that faculty members were "lenient" in assigning both grades and ratings. With respect to grades, faculty were very lenient. Mean FYAs range from 1.3 to 1.7 points above the theoretical midpoint of 2, which corresponds to a grade of C. However, faculty were substantially less lenient when they rated students. Mean ratings over all the scales ranged from .2 to .6 of a point above the scale midpoint of 3. The level of ratings was, however, a clear indication of a tendency toward leniency, because raters were asked to indicate students' abilities in relation to "other students they have known." A mean rating greater than 3.0 (the value indicating that a student was neither more nor less able than other students) would not be expected unless graduate students have become significantly more able recently or unless our sample of students did not fairly represent graduate students in the departments that participated in the study. In addition, the ratings appear to be less subject to restriction in range than is FYA. The standard deviations of the ratings were two to three times larger than those for FYA in Table 3.

The degree to which the six rating scales were independent was evaluated also. For each department the intercorrelations among the students' mean scores (averaged over raters) on each scale were obtained. Table 5 presents the median correlations among the scales for the 24 departments in the sample. The median correlations among the scales ranged from .76 to .87, suggesting that faculty did not differentiate students appreciably with respect to these aspects of reasoning. Interrater reliability was also examined. The median interrater correlation among students' mean scores (averaged over scales) was .40. (The computation of the traditional intraclass correlation coefficient was not undertaken because of the pattern of sparse data.)

In summary, preliminary analyses indicated that faculty were very lenient in their assignment of grades to graduate students and somewhat less lenient in their ratings of students. Furthermore, ratings were considerably less restricted in range then were grades. Unfortunately, we found little evidence that faculty were able to differentiate individuals with respect to the dimensions of reasoning identified in a previous study. Interrater agreement was modest, but the value of high interrater agreement in rating has been questioned because, for example, high agreement may reflect bias or halo effects (Saal, Downey, & Lahey, 1980), whereas low agreement may simply reflect different points of view among raters.

Exploring the Use of Adjusted Ratings

One major concern that guided subsequent analyses was the usefulness of FYA and faculty ratings of students as criteria for gauging the validity of GRE General Test scores. A second question was whether the GRE analytical measure is a better predictor of faculty ratings of students' analytical abilities than are the GRE verbal and quantitative measures. In these analyses, ratings were averaged over both raters and scales for each student to produce an overall mean rating of analytical ability for each individual. These analyses were exploratory in nature and relied, for the most part, on the comparison of zero-order correlations because of the instability of estimates of the validity of two or more predictors for small samples based on least squares regression. Methods exist to overcome this instability (Braun & Jones, 1985), but their application to the data gathered in this study was not straightforward, and therefore not attempted.

The ideal design for this study would have entailed having each faculty member provide ratings for exactly the same set of students, and our instructions to raters encouraged this ideal, which, unfortunately, was unattainable. Often there was little if any overlap among sets of students rated by different faculty. Because faculty probably differed with respect to the standards they applied when rating students, it was thought to be desirable to adjust for differences among faculty in the average levels of their ratings.

The planned method for accomplishing this was to use a general linear model analysis of variance program for one observation per cell with missing entries. In essence, missing values were imputed by estimating both an effect for each rater (to account for differing standards) and an effect for each student (to account for differing ability levels). Implementing this procedure with any confidence required at least some overlap among raters and ratees. Unfortunately, the data were quite sparse in many instances, and the method could not be applied with much confidence. It was possible, however, to make adjustments for some departments, in particular psychology departments, which tended more often than other departments to comply with our instructions.

Adjusted ratings were computed for each of seven psychology departments and then correlated with GRE scores, undergraduate grade point averages, and first-year graduate averages. These correlations were then compared with those based on unadjusted ratings. There were few consistent differences between the correlations for adjusted and unadjusted ratings, although there was a very slight tendency for higher correlations based on adjusted ratings. Adjustments were therefore not attempted for departments having even sparser data, and all analyses were based on unadjusted ratings. If we had been able to base our results on adjusted ratings, the predictability of ratings might have been slightly higher.

Faculty Ratings by Discipline

In the earlier study (Powers & Enright, 1987) that served as the basis for the development of faculty ratings, there were distinct differences among disciplines in the extent to which faculty perceived various analytical skills as important (Figure 1a). For example, in English departments argumentation skills were considered to be more important than the ability to analyze problems, whereas in computer science departments the opposite was true. The actual levels of ratings assigned in the different disciplines (Figure 1b) do not, however, correspond with the profiles obtained for perceptions of importance. Rather, the mean ratings assigned in each discipline showed quite flat profiles across the separate scales. Generally, ratings were no higher for one scale than for another. This result may reflect a lack of the discriminant validity of the individual scales. Alternatively, it may be a function of the wording of the rating scales. That is, faculty were instructed to provide ratings for each scale in relation to students they had known previously. With this instruction, we should not expect the level of ratings to vary among scales, because on average students should be neither more nor less able than previously known students.

Relationship between Ratings and Preadmission Measures

Table 6 shows for each discipline the correlations of analytical ratings (averaged over all scales) with GRE General Test scores and undergraduate grade average. Median correlations over departments are given, as well as correlations based on all students pooled over all departments within a discipline. The median and pooled correlations are not entirely consistent because of such factors as difference among departments in the numbers of students enrolled and in the average level of students' GRE scores.

Generally, undergraduate grades were less highly associated with analytical ratings than were GRE scores. There was, however, no particularly consistent tendency for ratings to relate more highly to GRE analytical scores than to verbal or quantitative scores.

Results did seem to vary somewhat by discipline. For example, ratings made by English faculty were more strongly related to GRE verbal scores than to other measures. However, correlations fluctuated from department to department, and few reliable trends could be discerned.

Faculty Ratings and First-Year Average as Criteria

To compare faculty ratings of students' analytical abilities and FYA as criteria, correlations of these measures with GRE scores and undergraduate grade point averages were calculated for each of the 24 participating departments. The median correlation between the two criteria themselves--FYA and faculty ratings--was .60 over all departments, indicating that faculty ratings and FYA have a common basis.

The median correlations between the four predictors and the two criteria are presented in Table 7. (Table B.1 in Appendix B presents these correlations for each department.) For comparison purposes, data reported by Willingham (1974), Burton and Turner (1983), and Schneider and Briel (1990) are included in Table 7. Little is known about the specific nature of the faculty ratings provided in these studies. However, most departments in Burton and Turner and Schneider and Briel probably rated student performance as "distinguished," "good," "adequate," or "unsatisfactory" with regard to departmental standards, since this is the scale that is mentioned in the Validity Study Service handbook. It is likely, in any event, that these ratings involved traits or accomplishments that were more general than the analytical abilities rated in the study reported here.

In the current study, the faculty ratings were on average predicted somewhat better than FYA by GRE scores, especially verbal scores, and somewhat worse by UGPA. On average, GRE quantitative scores and GRE analytical scores were only slightly more highly related to ratings than to FYA. These data can be compared to those reported by Willingham (1974) and Burton and Turner (1983) for the GRE verbal and quantitative measures. In each of these earlier studies, the correlation of GRE scores was generally higher with faculty ratings than with graduate FYA. However, in contrast to the study reported here, the Willingham and Burton and Turner studies found better prediction of faculty ratings than graduate FYA from undergraduate grade average. This was the most striking difference among these studies. In the present study, UGPA predicted graduate FYA much better than it predicted faculty ratings. In the other two studies, however, UGPA predicted faculty ratings slightly better than it predicted graduate FYA. This difference may reflect the fact that, in the current study, only one rating instrument, specifically designed to focus on a particular skill area, was used. However, the faculty rating instruments in the other studies varied among departments and focused on general performance.

The most recent data (Schneider & Briel, 1990), which are based on somewhat different samples of departments for first-year averages and faculty ratings, show somewhat different patterns of correlations. Faculty ratings appear to relate slightly less strongly on average with each predictor than do first-year grades.

Little evidence for the discriminant validity of the current version of the GRE analytical measure is evident in Table 7. The median correlation between the GRE verbal measure and the faculty ratings found in this study is slightly larger than the correlation between the GRE analytical measure and the faculty ratings. One reason for this is that reasoning contributes to performance on all three GRE measures. The rating instrument may have, in fact, focused faculty attention on students' verbal reasoning skills as well as their analytical skills.

When the correlations among the four predictive measures and the two criteria were ranked within each department in terms of size as in Table 8, the analytical measure fared much better. With respect to faculty ratings,

the analytical measure was the second best predictor for more than half the departments in the study and was either the first or second best predictor more often than any other measure. This contrasts with the role of the analytical measure in predicting FYA, for which it was only the third best predictor, perhaps because of the importance of UGPA in predicting FYA. UGPA, on the other hand, was seldom (4 times) the best or second best predictor of ratings, whereas it was most often (17 times) the best or next best predictor of FYA. The difference between FYA and ratings with respect to the ranking of predictors (as either best or next best vs. worst or next worst) was statistically significant both for GRE analytical scores, $X^2(1) = 6.80$, $p < .01$, and for undergraduate grade average, $X^2(1) = 12.19$, $p < .01$. Taken together, Tables 7 and 8 demonstrate that the predictive power of the several predictors varies with each criterion. This in turn suggests that these criteria may be tapping different aspects of graduate student success, with analytical ratings relatively more reflective of analytical ability (as measured by the GRE analytical measure) and first-year average relatively more reflective of prior academic achievement (as indexed by undergraduate grade average).

Role of Faculty Familiarity with Students

Previous research has suggested that the nature of raters' contact with ratees may affect the quality of ratings, hence their relationship to other variables (e.g., Freeberg, 1969; Landy & Farr, 1980). In providing their ratings in this study, faculty were also asked to describe how much opportunity (significantly less, slightly less, neither more nor less, slightly more, or significantly more compared with other students) they had to observe/judge the extent to which the students they rated possessed the kinds of analytical skills of interest.

The role of familiarity was assessed by first regressing analytical ratings (averaged over all scales and raters) on GRE analytical scores. Ratings were first converted to z-scores within each department (to adjust for possibly different standards) and then pooled across all departments for each discipline. Next, a variable reflecting degree of familiarity with students was added, and the contribution to the multiple R^2 was assessed. Finally, a product variable (interaction of GRE analytical score x familiarity) was added, and its contribution was assessed as an indication of the degree to which the prediction of ratings from GRE analytical scores was moderated by the degree of familiarity with students.

The correlations of GRE analytical scores with faculty rating were .14, .32, .26, .37, .31, and .28 for all students pooled across departments for chemistry, computer science, education, engineering, English, and psychology, respectively. In none of the disciplines did the interaction term contribute significantly to the prediction of ratings beyond the contribution of GRE analytical scores and familiarity. This suggests that the relationship of ratings to GRE analytical scores does not depend on the degree to which faculty are acquainted with students. (All ratings in the study were made

only of students with whom faculty had made at least some contact. Furthermore, the mean ratings on the familiarity scale were greater than 3.0, the scale midpoint, for all but one department and greater than 3.5 for 10 of the 24 departments.) In three disciplines, however--computer science, engineering, and English--the level of ratings was significantly related ($p < .05$) to familiarity with students, with higher ratings given to students with whom faculty were better acquainted, suggesting perhaps that irrelevant social contact may have played a part in the ratings. Another possible interpretation is that faculty may have had more relevant contact with more able students (e.g., as research or teaching assistants) and therefore greater opportunity to observe their performance.

Usefulness of a Composite Criterion

Because faculty ratings and FYA seemed to be tapping different aspects of accomplishment, it was thought that combining faculty ratings and FYA into a joint criterion might increase the validities of the predictors (see also Wild, Swinton, & Brown, in preparation). FYA and mean faculty ratings were converted to z-scores within each department and added together to obtain a composite score for each student (cf. Wild, Swinton, & Brown). Zero-order correlations were computed between the predictors and the standardized criteria, individually and in combination. In these analyses, only departments with data for at least 10 students on all four predictors and both criterion measures were included. This resulted in the exclusion of two English departments from the analysis. The number of students included in the analyses for the remaining departments varied from 11 to 33.

The results of these analyses are summarized in Table 9. For the GRE measures, the median correlations were higher for the prediction of the composite than for either criterion alone. However, UGPA was more highly related to FYA than to the joint criterion.

One implication of these results is that the use of faculty ratings either alone or in combination with FYA may increase the contribution of GRE scores and reduce that of UGPA to the prediction of graduate success. The use of a composite criterion that incorporates information collected from different perspectives may present a more balanced picture of success in graduate school and relate differentially to various predictors.

Summary and Discussion

The aim of the study reported here was to develop a criterion of success in graduate education as an alternative to the traditionally used criterion of first-year graduate grade average. In particular, the objective was to evaluate and explore the use of faculty ratings of students' analytical skills or abilities as a potential criterion against which to gauge the validity of

the current GRE analytical measure. More important, however, the intention was to make available a suitable criterion that would facilitate the assessment of progress toward an improved and more defensible measure of analytical ability than is now offered.

Six rating scales were developed on the basis of previous empirical research that gathered graduate faculty perceptions of the importance for successful graduate study of a wide variety of analytical, reasoning, or thinking skills. The particular features of the scales were chosen according to available research on ratings and on the basis of suggestions made by graduate faculty regarding the feasibility of several different procedures.

Representative samples of graduate departments in each of six disciplines--chemistry, computer science, education, engineering, English, and psychology--were invited to participate in the study. Although the response was less than hoped for, the invitation yielded a total sample of 24 graduate departments in which 132 faculty members provided ratings for a total of 623 graduate students. These departments and students did not constitute a random sample, but they did provide a relatively good cross-section of graduate departments in the six disciplines.

Flat profiles of ratings across scales over disciplines and high correlations among the six scales suggested little if any discriminant validity of the individual scales: each scale seemed to reflect a single, more general trait. Most of the analyses were based therefore on the total of ratings over all scales to assess the validity of the ratings as an indicator of this more general analytical trait.

The ratings had moderately good reliability. Although interrater agreement was relatively modest, correlations of ratings with other variables suggested an adequate level of reliability. Also, the very high correlations among scales suggested either substantial reliability or the existence of a significant halo effect (although the way in which ratings were collected was thought to have minimized the likelihood that a particular student would be placed at the same level on different scales).

Faculty tended to be somewhat lenient in making the ratings, even though they were to be used only for research purposes, not for student evaluation. However, ratings exhibited substantially greater variation and significantly less leniency than did first-year averages.

Ratings appeared to be strongly related to an alternative indication of student success. A median correlation of .60 with first-year graduate average suggested that faculty ratings and graduate grades both reflect academic success, but may not be completely interchangeable, even though this correlation is quite high in relation to the likely reliability of the two indicators.

Several alternative analyses yielded mixed results with respect to the validity of faculty ratings of students' analytical abilities. Each of the

three GRE General Test scores, especially verbal scores, correlated slightly higher on average with ratings than with first-year graduate averages. Verbal scores were more highly associated with ratings than were analytical scores, suggesting perhaps that judgments of analytical ability were based at least in part on students' verbal reasoning skills. Undergraduate grade point averages, on the other hand, correlated better with first-year averages than with ratings, suggesting that ratings may reflect an aspect of achievement that is not captured by grades. When undergraduate grade average and GRE scores were ranked with respect to their predictive effectiveness in each department, GRE analytical scores were significantly more often one of the two best predictors of faculty ratings than of first-year averages. Undergraduate grades, on the other hand, were significantly more often one of the two best predictors of first-year graduate grades than of ratings. GRE verbal and quantitative scores tended to be the best predictors about equally often for ratings and grades. These patterns reinforce the notion that graduate grades and faculty ratings reflect different aspects of accomplishment.

The potential of the analytical ratings as a criterion was also apparent from the increased prediction when ratings and first-year graduate averages were combined to form a composite criterion. The prediction of this composite from GRE scores was on average better than the prediction of either first-year averages or ratings individually, possibly because this composite may be more reliable.

On balance, the faculty ratings of students' analytical or reasoning skills that were developed in this study appear to have some modest potential as an alternative criterion of success against which to assess progress toward an improved measure of analytical ability. The use of these ratings is not completely problem-free, however. Difficulties arose in securing faculty interest in completing the ratings, quite possibly because the ratings were difficult ones to make. Instead of the relatively global judgments that are typically sought in rating studies, the scales developed here required faculty to think of quite specific manifestations of the traits of interest. Revising the scales to make them more global in focus might improve their usability. In addition, any future use or study of these or similar scales should probably focus on securing the cooperation of a smaller number of departments and faculty who are most interested in the possible use of such ratings. The degree to which departments and faculty are committed to providing usable ratings may be an important factor in the extent to which ratings prove to be valid indications of students' abilities.

A problem that may be more difficult to overcome is the apparent general lack of faculty acquaintance with many students. In attempting to assess reliability and to ensure the comparability of ratings, we strove to obtain ratings by several faculty of common sets of students. Typically, however, faculty were unable to comply with this request because of their very unequal familiarity with students. In any future studies that employ these rating scales, extra effort should probably be directed toward identifying departments in which faculty have close contact with students.

With respect to the validity of the ratings, the most troublesome aspect was the lack of any consistent evidence of discriminant validity. Instead of correlating higher with GRE analytical scores than with verbal or quantitative scores, ratings were on average more highly related to GRE verbal scores than to analytical scores, suggesting the possibility that students' verbal reasoning skills were influencing faculty ratings of students' analytical skills. These patterns of correlations may, however, also reflect the lack of discriminant validity of the current version of the analytical measure.

Because little information would be lost by combining ratings from the six scales, the rating instrument could be streamlined to facilitate use. For example, one or two of the most appropriate scales could be selected and combined for each discipline, with ratings of students in English departments, for instance, emphasizing argumentation skills and those in computer science emphasizing problem analysis. Alternatively, although the separate scales seem to reflect mainly one dimension, several could be retained to ensure an adequate level of reliability.

In conclusion, this study has provided some modest progress toward the development of an alternative criterion of success in graduate school and, more specifically, a criterion that deserves further attention in future research on improving the current GRE analytical measure. In addition, the rating scales may prove to be a useful addition to the GRE Validity Study Service. Although the scales may not enjoy widespread, routine use, some departments may appreciate their availability. First, however, further developmental research might be undertaken to refine the scales as they now exist and to gather more conclusive evidence of their validity.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Anastasi, A. (1979). Fields of applied psychology (2nd edition). New York: McGraw-Hill Book Company.

Braun, H. I., & Jones, D. H. (1985). Use of Empirical Bayes method in the study of the validity of academic predictors of graduate school performance. (GRE Board Professional Report GREB No. 79-13P, ETS Research Report 84-34). Princeton, NJ: Educational Testing Service.

Burton, N., & Turner, N. (1983). Effectiveness of Graduate Record Examinations for predicting first-year grades: 1981-82 summary report of the Graduate Record Examinations validity Study Service. Princeton, NJ: Educational Testing Service.

Carlson, A. B., Reilly, R. R., Mahoney, M. H., & Casserly, P. L. (1976). The development and pilot testing of criterion rating scales (GRE Board Professional Report No. 73-1P). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1988). A summary of data collected from Graduate Record Examinations test takers during 1986-87 (Data Summary Report #12). Princeton, NJ: Educational Testing Service.

Freeberg, N. E. (1969). Relevance of rater-ratee acquaintance in the validity and reliability of ratings. Journal of Applied Psychology, 53, 518-524.

Graduate Record Examinations Board and Council of Graduate Schools in the United States. (1983). Directory of graduate programs: 1984 and 1985 (4 volumes). Princeton, NJ: Educational Testing Service.

Harari, O., & Zedeck, S. (1973). Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology, 58, 261-265.

Hartnett, R., & Willingham, W. W. (1979). The criterion problem: What measures of success in graduate education? (GRE Research Report No. 77-04). Princeton, NJ: Educational Testing Service.

Jenkins, G. D., & Taber, T. A. (1977). Monte Carlo study of factors affecting three indices of composite scale reliability. Journal of Applied Psychology, 62, 392-398.

Kingston, N. M. (1985, April). Incremental validity of the GRE analytical ability measure for predicting graduate first-year graduate point average. Paper presented at the annual meeting of the American Research Association, Chicago.

Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

Landy, F. J., & Guion, R. M. (1970). Development of scales for the measurement of work motivation. Organizational Behavior and Human Performance, 5, 93-103.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability. A Monte Carlo approach. Journal of Applied Psychology, 60, 10-13.

National Center for Educational Statistics. (1984). Institutional characteristics of colleges and universities: 1983-84 (IC Survey, HEGIS XVIII machine-readable data file). Washington, DC: Department of Education.

Powers, D. E., & Enright, M. K. (1986). Analytical reasoning skills involved in graduate study: Perceptions of faculty in six fields (GRE Board Professional Report No. 83-23P, ETS Research Report 86-43). Princeton, NJ: Educational Testing Service.

Powers, D. E., & Enright, M. K. (1987). Analytical reasoning skills in graduate study: Perceptions of faculty in six fields. Journal of Higher Education, 58, 658-682.

Powers, D. E., & Swinton, S. S. (1981). Extending the measurement of graduate admission abilities beyond the verbal and quantitative domains. Applied Psychological Measurement, 5, 141-158.

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. Journal of Educataional Psychology, 76, 266-278.

Rock, D. A., Werts, C., & Grandy, J. (1982). Construct validity of the GRE Aptitude Test across populations--An empirical confirmatory study (GRE Board Professional Report No. 78-1P, ETS Research Report 81-57). Princeton, NJ: Educational Testing Service.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.

Schaeffer, G. A., & Kingston, N. M. (1988). Strength of the analytical factor of the GRE General Test in several subgroups: A full-information factor analysis approach (GRE Board Professional Report No. 86-7P, ETS Research Report 88-5). Princeton, NJ: Educational Testing Service.

Schneider, L. M., & Briel, J. B. (1990). Validity of the GRE: 1988-89 summary report. Princeton, NJ: Educational Testing Service.

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago, Rand McNally College Publishing Company.

Stricker, L. J., & Rock, D. A. (1985). Factor structure of the GRE General Test for older examinees: Implications for construct validity (GRE Board Research Report No. 83-10R, ETS Research Report 85-9). Princeton, NJ: Educational Testing Service.

Swinton, S. S., & Powers, D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. Journal of Educational Psychology, 75, 104-115.

Swinton, S., Wild, C., & Wallmark, M. (1983). Investigation of practice effects on item types in the Graduate Record Examinations Aptitude Test (GRE Report No. 80-1cP, ETS RR 82-56). Princeton, NJ: Educational Testing Service.

Wild, C. L., & Swinton, S. S., & Brown, F. G. (in preparation). Validity of the Graduate Record Examinations (GRE) using faculty rating and grade point average (Report being submitted to the GREB Research Committee in September). Princeton, NJ: Educational Testing Service.

Wild, C. L., Swinton, S. S., & Wallmark, M. M. (1982). Research leading to the revision of the format of the Graduate Record Examinations Aptitude Test in October 1981 (GRE Board Professional Report No. 80-1bP, ETS Research Report 82-55). Princeton, NJ: Educational Testing Service.

Willingham, W. W. (1974). Predicting success in graduate education. Science, 183, 273-278.

Wilson, K. M. (1982). A study of the validity of the restructured GRE Aptitude Test for predicting first-year performance in graduate study (GRE Board Research Report No. 78-6R, ETS Research Report 82-34). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1985). The relationship of GRE General Test item-type part scores to undergraduate grades (GRE Professional Report GREB No. 81-22P, ETS RR 84-38). Princeton, NJ: Educational Testing Service.

Table 1

Dimensions Underlying Graduate Faculty Perceptions

Of Analytical Ability

	Skills
I	Argumentation: Critical thinking related to analyzing and evaluating arguments
II	Explanation: Critical thinking related to drawing inferences and developing conclusions
III	Problem Analysis: Ability to define and set up problems
IV	Induction: Ability to reason inductively
V	Generation of Alternatives: Ability to generate alternative explanations/hypotheses
	Incidents
I	Critical Facility: Reliance on evidence, and "hard data" versus emotional appeal and anecdotal information
II	Inability to generate alternatives
III	Practical judgment/common sense

Table 2

GRE Rating Scales: Analytical Ability

1.	Critical Thinking: Argumentation Ability to understand, analyze, and evaluate arguments
2.	Critical Thinking: Drawing Inferences and Developing Conclusions Ability to construct sound inferences and conclusions
3.	Defining Problems Ability to define and set up problems
4.	Inductive Reasoning Ability to reason from specific instances to more general principles
5.	Generating Alternatives Ability to generate alternative explanations or hypotheses
6.	Analytical Style Inclination toward analytical or critical thinking

Table 3

Mean GRE Scores, UGPA, and Graduate FYA for Participating Departments by Discipline

Discipline (Nd)*		GRE V	GRE Q	GRE A	UGPA	FYA
Chemistry (2)	N**	53	53	43	50	53
	Mean	458	675	551	3.2	3.3
	SD	130	85	127	.6	.4
Computer Science (3)	N	71	71	67	67	75
	Mean	529	682	616	3.2	3.5
	SD	143	80	116	.4	.4
Education (4)	N	115	115	106	117	123
	Mean	473	494	507	3.2	3.7
	SD	96	102	112	.5	.3
Engineering (4)	N	91	91	90	83	95
	Mean	504	711	607	3.3	3.6
	SD	145	71	129	.6	.3
English (4)	N	85	85	77	68	86
	Mean	671	568	595	3.4	3.6
	SD	74	97	106	.5	.4
Psychology (7)	N	150	150	150	147	150
	Mean	618	611	614	3.5	3.7
	SD	85	82	93	.3	.3

* Nd = numbers of departments participating
** N = number of students in each discipline

Table 4

Mean Ratings on Scales for Participating Departments by Discipline

Discipline		Argumentation	Inferences & Conclusions	Problem Analysis	Inductive Reasoning	Generating Alternatives	Analytical Style	Mean of Six Scales
Chemistry	N	53	51	36	50	33	51	53
	Mean	3.2	3.1	3.7	3.1	3.4	3.1	3.2
	SD	1.2	1.1	1.1	1.2	1.1	1.3	1.1
Computer Science	N	75	75	73	71	73	74	75
	Mean	3.3	3.4	3.4	3.3	3.4	3.4	3.3
	SD	1.0	1.0	0.9	1.0	0.9	1.0	0.9
Education	N	122	122	122	122	122	122	122
	Mean	3.6	3.7	3.6	3.5	3.5	3.5	3.6
	SD	0.9	0.9	0.9	0.9	1.0	0.9	0.8
Engineering	N	94	94	94	93	93	94	94
	Mean	3.4	3.4	3.3	3.3	3.4	3.4	3.4
	SD	0.9	0.8	0.8	0.9	0.8	0.8	0.8
English	N	85	84	84	83	84	84	85
	Mean	3.6	3.6	3.5	3.5	3.5	3.6	3.5
	SD	0.9	0.8	0.8	0.8	0.8	0.9	0.8
Psychology	N	150	150	150	149	149	150	150
	Mean	3.4	3.4	3.4	3.4	3.4	3.5	3.4
	SD	0.9	0.9	0.9	0.9	0.8	0.9	0.8

Table 5

Median Intercorrelations Among Rating Scales for All Departments

	Argumentation	Inferences & Conclusions	Defining Problems	Inductive Reasoning	Generating Alternatives
1. Argumentation					
2. Inferences & Conclusions	.87				
3. Defining Problems	.81	.79			
4. Inductive Reasoning	.82	.84	.79		
5. Generating Alternatives	.76	.78	.79	.81	
6. Analytical Style	.82	.81	.80	.81	.77

Table 6

Average Correlations of GRE Scores and Undergraduate Grade Point Average with Mean Analytical Rating by Discipline

Discipline		Variable			
		GRE V	GRE Q	GRE A	UGPA
Chemistry	Median	-.07	.38	.21	.07
	Pooled	.04	.33	.24	-.06
Computer Science	Median	.42	.18	.24	.10
	Pooled	.19	.17	.32	.28
Education	Median	.33	.26	.38	.26
	Pooled	.34	.25	.28	.24
Engineering	Median	.24	.24	.36	.28
	Pooled	.32	.36	.33	.27
English	Median	.47	.11	.30	.21
	Pooled	.53	.18	.32	.11
Psychology	Median	.39	.41	.43	.18
	Pooled	.33	.31	.27	.15

Note. Median correlations are computed over departments. Pooled correlations are based on all students pooled over departments.

Table 7

Correlations of GRE Measures and UGPA with Faculty Rating and FYA

Current Study: Median Correlations for 24 Departments

Criterion	GRE V	GRE Q	GRE A	UGPA
Faculty Ratings	.39	.30	.34	.18
FYA	.25	.27	.30	.35

Willingham (1974): Median Correlations (Number of Departments) Reported in 43 Studies from 1952 to 1972

Criterion	GRE V	GRE Q	GRE A	UGPA
Faculty Ratings	.31 (27)	.27 (25)	NA	.37 (15)
FYA	.24 (46)	.23 (43)	NA	.31 (26)

Burton & Turner (1983): Size-Adjusted Average Correlations for 20 Departments: Studies through June 1982

Criterion	GRE V	GRE Q	GRE A	UGPA
Faculty Ratings	.22	.24	NA	.35
FYA	.14	.16	NA	.31

Schneider & Briel (1990): Size-Adjusted Average Correlations (Number of Departments) (September 1984-September 1988)

Criterion	GRE V	GRE Q	GRE A	UGPA
Faculty Ratings (89)	.25	.25	.21	.31
FYA (606)	.29	.28	.26	.34

Table 8

Rank of Predictive Measures' Correlations with Faculty Ratings and FYA Summarized Over 24 Departments

Prediction of Faculty Ratings

Measures	First	Second	Third	Fourth
GRE V	10	3	6	5
GRE Q	7	6	4	7
GRE A	3	15	4	2
UGPA	4	0	10	10

Prediction of FYA

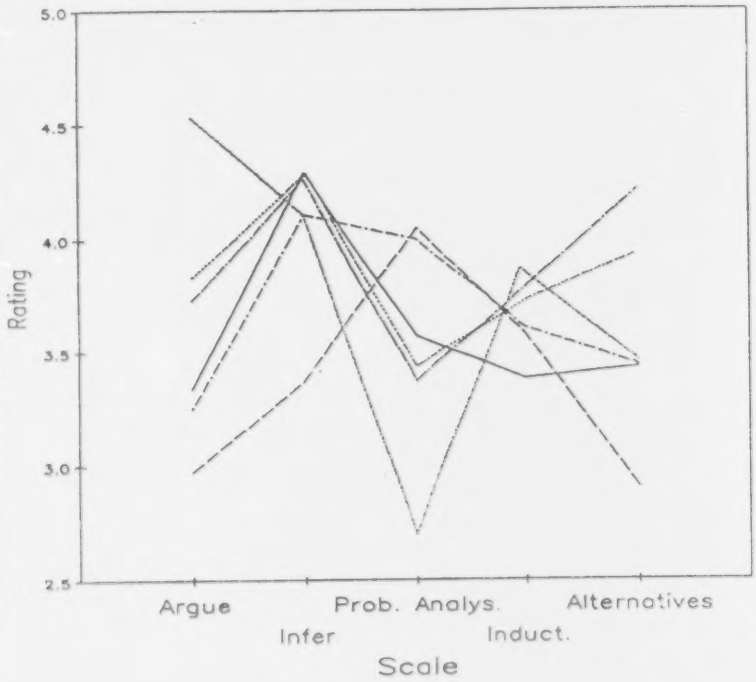
Measures	First	Second	Third	Fourth
GRE V	6	5	4	9
GRE Q	6	6	4	8
GRE A	1	7	13	3
UGPA	11	6	3	4

Table 9

Median Correlations of Four Predictors with Standardized Criteria for 22 Departments

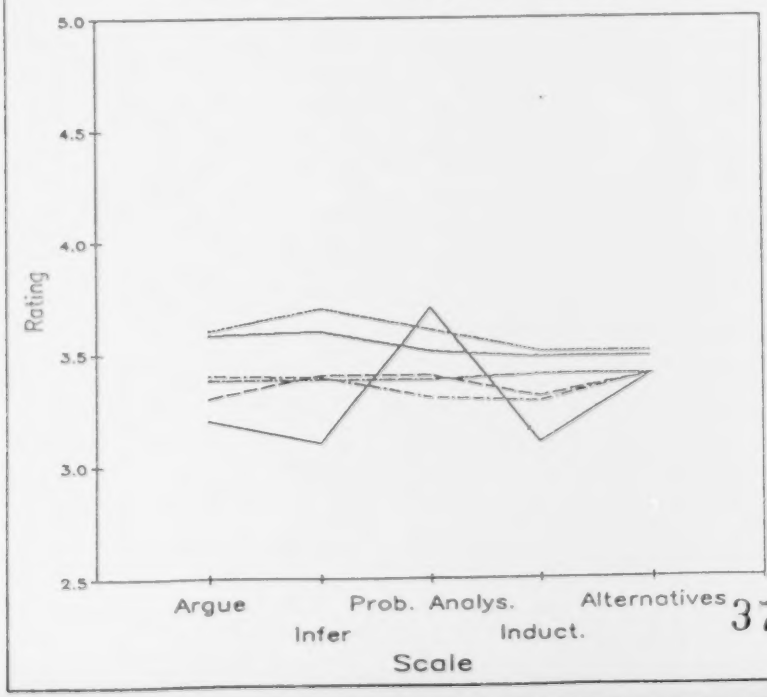
	Predictors			
Criterion	GRE V	GRE Q	GRE A	UGPA
Faculty Ratings	.37	.31	.35	.17
FYA	.27	.32	.33	.34
Faculty Rating & FYA Composite	.40	.36	.37	.28

Figure 1a. Mean Ratings of Importance of Reasoning Skills by Discipline



Source: Powers, D. E., & Enright, M. K. (1987). Analytical reasoning skills in graduate study: Perceptions of faculty in six fields. Journal of Higher Education, 58, 658-682.

Figure 1b. Mean Faculty Ratings of Students' Reasoning Skills by Discipline



Appendix A

Copyright © 1987 by Educational Testing Service. All rights reserved.

GRE RATING SCALES:

ANALYTICAL ABILITY

ANALYTICAL ABILITY RATING SCALES

The attached rating scales, intended for research purposes only, have been developed under the auspices of the Graduate Record Examinations (GRE) Board. They have been designed to provide further information about the validity of GRE analytical ability scores, the newest GRE scores, which are now reported with GRE verbal and quantitative scores. The scales are based on information obtained in a nationwide survey of graduate faculty, which identified several dimensions of analytical skills thought to be especially important for success in graduate school.

Six analytical traits are described on the attached forms. Names of the students to be rated are listed on the left. When making your ratings, think of each student in terms of the descriptions provided on each page, and write the number that best describes each student. Some students will undoubtedly be more difficult to rate than others, but please attempt to rate each student with whom you've had some contact. Some dimensions may be less important in your field than in others, but we would like you to rate students on each scale if possible. The final scale asks about your familiarity with each student.

As you proceed through the different ratings, think of each student only in terms of the particular trait being rated. Please do not refer back to previous ratings, but instead make each rating as independently as possible from the previous ones. We have tried to design the forms so that you will be less likely to be influenced by previous ratings.

Please make your ratings in reference to "graduate students you have known," not just in terms of the current students. In doing so, try to use the full 5-point range. Don't be reluctant, as many raters often are, to use the extreme values of the scale, and try to resist, as many raters fail to do, overusing the middle values of the scale. For example, you might try to use the middle value for no more than a third of the students.

As you think about students, we encourage you to remember both positive and negative instances of performances. One strategy that you might wish to use is to read the list of students whom you will rate, and to become generally familiar with each trait. Then, put the scales aside for a period of a week or so, during which you may observe examples of behaviors that exemplify the traits to be rated. After you have made your ratings, please return the forms to the study coordinator at your department, who will forward them to the project directors. Your ratings will be treated confidentially, and results will be reported only in terms of relationships among test scores, grades, and ratings. No individual students or faculty will be identified, and all identifying information will be deleted from files when data matching is complete. We greatly appreciate your cooperation and effort in completing these forms.

Thank you.

Rating Scale A

Name of Rater _____

Name of Student _____

Write the number that best describes each student

CRITICAL THINKING: ARGUMENTATION

Description: Ability to understand, analyze, and evaluate arguments

_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )

Generally, a student possessing this trait:

- Tends to know what kind of evidence will support or refute a hypothesis
- Can typically recognize the central argument or thesis in a work
- Can usually identify both stated and unstated assumptions in an argument
- Is likely to recognize fallacies and logical contradictions in arguments
- Can elaborate an argument and develop its implications

SCALE

Compared with other students I have known, this student is:

1 = Significantly less able

2 = Slightly less able

3 = Neither more nor less able

4 = Slightly more able

5 = Significantly more able

0 = I've had no opportunity to observe this student with respect to this skill.

Rating Scale B

Name of Rater

_____

Name of Student

Write the number that best describes each student

( )

( )

( )

Rating Scale C

Name of Rater

_____

Name of Student _____

Write the number that best describes each student

_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )

DEFINING PROBLEMS

Description: Ability to define and set up problems

Generally, a student possessing this trait:

- Can usually break down complex problems into simpler ones
- Can typically identify most of the variables or factors involved in a problem
- Can, when appropriate, set up a formal model for a problem under consideration
- Is usually able to translate graphs or other symbolic statements into words and vice versa
- Can develop operational (or very precise) definitions of concepts

SCALE

Compared with other students I have known, this student is:

1 = Significantly less able

2 = Slightly less able

3 = Neither more nor less able

4 = Slightly more able

5 = Significantly more able

0 = I've had no opportunity to observe this student with respect to this skill.

Rating Scale D

Name of Rater

_____

Name of Student

_____

Write the number that best describes each student

INDUCTIVE REASONING

Description: Ability to reason from specific instances to more general principles

_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )
_____ ( )

Generally, a student possessing this trait:

- Shows ability to derive general or abstract principles from disparate facts or cases
- Can often solve problems in situations in which all the necessary information is not known
- Typically is able to recognize structural similarities between one type of problem, theory, or idea and another
- Is capable of synthesizing two different positions into a third one

SCALE

Compared with other students I have known, this student is:

1 = Significantly less able

2 = Slightly less able

3 = Neither more nor less able

4 = Slightly more able

5 = Significantly more able

0 = I've had no opportunity to observe this student with respect to this skill.

Rating Scale E

Name of Rater

_____

Write the number that best describes each student

Name of Student

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

_____ ( )

GENERATING ALTERNATIVES

Description: Ability to generate alternative explanations or hypotheses

Generally, a student possessing this trait:

- Can find alternative explanations for observations
- Is able to generate alternative hypotheses
- Is inclined to search for counterexamples to test the validity of an argument or explanation
- Typically recognizes two or more sides of an issue
- Can often generate new questions or experiments to extend or support an interpretation

SCALE

Compared with other students I have known, this student is:

1 = Significantly less able

2 = Slightly less able

3 = Neither more nor less able

4 = Slightly more able

5 = Significantly more able

0 = I've had no opportunity to observe this student with respect to this skill.

Rating Scale F

Name of Rater

_____

Name of Student

_____

Write the number that best describes each student

ANALYTICAL STYLE

Description: Inclination toward analytical or critical thinking

_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )

A student possessing this trait:

- Is unlikely to accept assumptions without questioning them
- Usually avoids making generalizations from insufficient evidence
- Can integrate and synthesize ideas
- Pays attention to important details
- Is able to detect patterns and to generalize from them

SCALE

Compared with other students I have known, this student is:

1 = Significantly less able

2 = Slightly less able

3 = Neither more nor less able

4 = Slightly more able

5 = Significantly more able

0 = I've had no opportunity to observe this student with respect to this skill.

Rating Scale G

Name of Rater

_____

Name of Student

_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )
_____	( )

Write the number that best describes each student

This scale is designed to determine the degree to which you are acquainted with each student. Please indicate for each student how much opportunity you have had to observe or judge (either positively or negatively) the kinds of traits described previously.

SCALE

Compared with most students, for this student I've had:

1 = Significantly less opportunity

2 = Slightly less opportunity

3 = Neither more or less opportunity

4 = Slightly more opportunity

5 = Significantly more opportunity

0 = I have had no contact with this student

Your cooperation and effort in completing these forms is greatly appreciated.

Appendix B

Table B.1

Correlations of GRE Measures and UGPA with
Mean Faculty Rating and FYA for All Departments

		With Mean Faculty Rating				With FYA			
Department		GREV	GREQ	GREA	UGPA	GREV	GREQ	GREA	UGPA
Chemistry	1	.07	.38	.09	.02	-.12	.21	.00	.10
	2	-.22	.39	.33	.12	.04	.61	.58	-.14
Computer Science	1	.24	.18	.13	.10	.11	.13	.24	.36
	2	.49	-.21	.24	.03	.20	.45	.35	.61
	3	.42	.78	.70	.55	.50	.75	.67	.48
Education	1	.16	.05	-.29	.04	.57	.14	.45	.43
	2	.39	.10	.39	.21	.37	.25	.33	.29
	3	.43	.41	.46	.31	.36	.26	.31	.39
	4	.27	.51	.36	.32	.49	.13	.42	.60
Engineering	1	.51	.43	.40	.10	.25	.41	.21	.21
	2	-.10	.29	.36	.17	.00	.51	.31	.34
	3	.09	.05	.12	.40	.01	.29	.25	.48
	4	.39	.19	.35	.48	-.16	.12	.09	.29
English	1	.30	.11	.33	.79	.31	.27	.28	.61
	2	.50	.09	-.47	-.57	.15	-.46	-.88	-.38
	3	.44	.12	.28	.25	.33	-.12	.11	.13
	4	.53	.31	.46	.17	.05	.26	.08	.39
Psychology	1	.07	.42	.43	.12	.06	.28	.25	.79
	2	.29	.23	-.05	.30	.46	.42	.17	.43
	3	.18	.41	.20	.14	.25	.07	.41	.27
	4	.39	.56	.54	.18	.63	.79	.64	.26
	5	.74	.32	.56	.18	.72	.31	.56	.26
	6	.60	.01	.09	.08	.28	.03	.18	.34
	7	.40	.64	.51	.34	.23	.46	.34	.46

54020-01160 • Y11M.7 • 209060